2003.12.02

Informal 小解説

例題で考える
一般化線形混合モデルの導入と計算

— 「個性」のモデリング? —

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2003/

講釈: 久保拓弥　kubo@ees.hokudai.ac.jp

今日のハナシ: 二値データを混合モデルで

1. **二値データ** (あり vs なし) **と一般化線形モデル**
2. **このとき，個体ごとに「個性」があったら**

   ここでいう「個性」とは以下のどれか，もしくはその組み合わせ:

   - 個体ごとに固有な性質のうち測定されなかったもの (遺伝子型など)
   - 個体のおかれている微環境のうち測定されなかったもの (土壌の状態など)

   このような「個性」のことを「ランダム効果」と呼んだりする—
   このような状況は混合モデルで取り扱う

3. **乱数による数値例から考える**

   答えがわかってる具体例について R の `glmmML()` を適用してみる

一般化線形モデル (generalized linear model, GLM)

- **指数関数族に属する確率分布あれこれ (正規分布，二項分布，ポアソン分布，$\cdots$) で説明されるばらつきのデータに適用できる**
- link **関数を指定できる**
- **独立変数は何でもよい: 連続変数，名義変数，順序変数**
- **パラメーターは線形に結合していなくてはならない (線形モデル)**

$$\mathrm{link}(\mu(\mathbf{x})) = \beta_0 \cdot 1 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \sum_i \beta_i x_i$$

統計ソフトウェア `R` **では** `glm()` **関数で簡単に推定計算ができる**

統計ソフトウェア R (M1 授業で使った宣伝)

http://www.r-project.org/

- いろいろな OS で使える freeware
- 使いたい機能が充実している
- 作図機能も強力
- S 言語によるプログラミング可能
- よい教科書が出版されつつある
  - 「R による統計解析の基礎」 中澤港 (2003)
  - “Introductory Statistics with R” P. Dalgaard (2002)
  - “Computational Statistics” M. Crawley (2002)
  - ネット上のあちこち

R の glm() で使える統計モデル

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	rbinom()	glm(family = binomial)
	二項分布	rbinom()	glm(family = binomial)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb()
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- 上記のさまざまな確率分布といくつかの link 関数(すなわち平均値の関数型) を組み合わせることができる
- さらに stepAIC() (in MASS library) と組み合わせると AIC によるモデル選択ができる

確率分布を推定する方法たちの階層性

なぜ一般化線形モデルを? — より広い世界を統一的に扱えるから

[**最尤推定法**で扱えるモデル]
何でもいいから確率分布があるモデル
一般化線形混合モデルなどなど

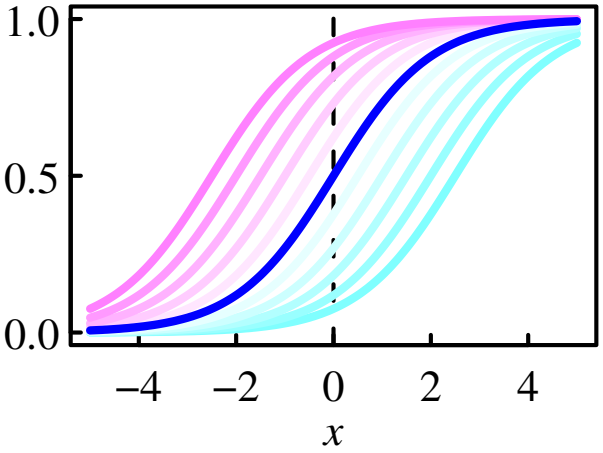
[**一般化線形モデル** (GLM)]
指数関数族の確率分布 + 線形モデル
ロジスティック回帰，ポアソン回帰などなど

[**最小二乗法**的に考えるモデル]
等分散正規分布 + 線形モデル
直線回帰，いわゆる「分散分析」などなど

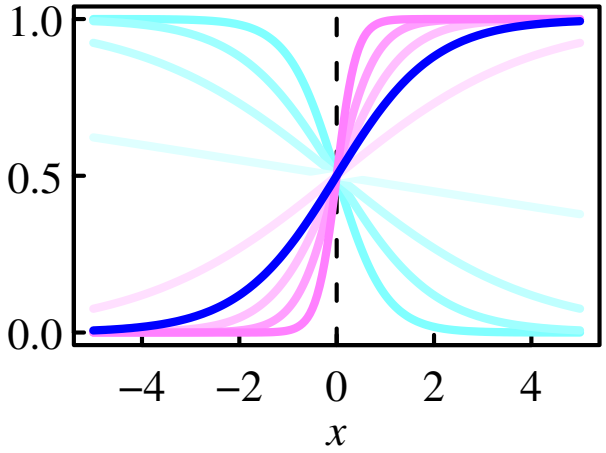
ついでに復習: ロジスティック曲線のカタチ

$$p(x) = \frac{1}{1 + \exp(-(a + bx))}$$

a だけ変化させる



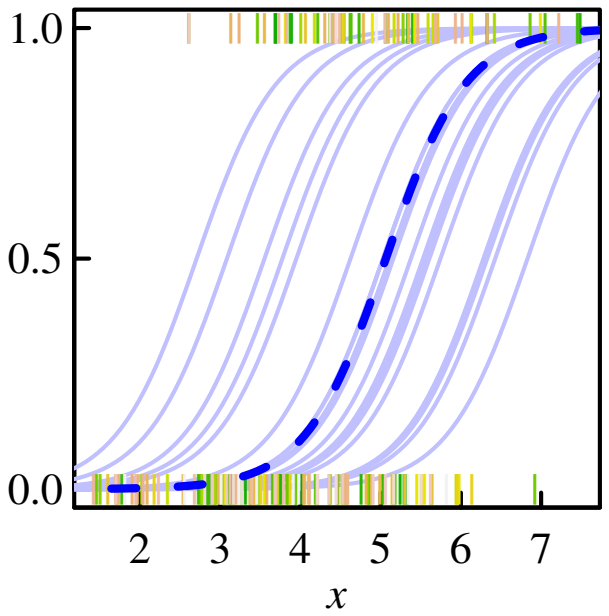
b だけ変化させる



具体的な数値例: サイズで変わる繁殖確率

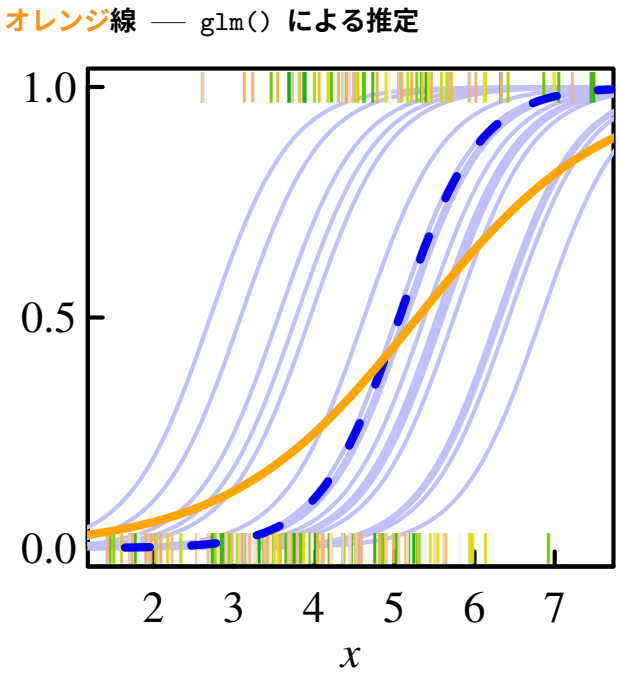
これは架空の草本　**繁殖確率**　$p(x) = \dfrac{1}{1+\exp(-(a+bx))}$

- 地上部のサイズ x によって繁殖する確率 $p(x)$ が変わる
- サンプル数 20 個体
- 各個体ごとの観測は 10 年
- 各個体・各年完全独立
- サイズ x は各年ごとに完全独立
- サイズ依存性である b は個体間で共通 (推定したい主な対象)



- 個体ごとにパラメーター a が異なる (個体ごとに固定した値)
  — そして a はある確率分布から得られた乱数である

「個性」無視して推定してみると



左の推定は R の glm()(混合ではない一般化線形モデルの計算) によるもの

- **「傾き」がかならず小さく偏って推定される**
- **Overdispersion (過分散) が発生している**

生態学の観測データにみられる「ばらつき」には,「個体差」(あるいは観測されてない微環境の違いなど) を無視できない場合があるかも

混合 (効果) モデル (mixed effects model)

Fixed + Random → Mixed effects

M. Crawley **先生本** "Computational Statistics" (2002) **によると，混合モデルを使ったほうが良さそうな生態学的な事例としては以下のような状況が考えられる**:

- **近傍個体間の空間的な自己相関**
- **同じ個体に対するくり返し計測にあらわれる時間的自己相関**
- **野外実験におけるブロック間の平均応答の差位**
- **くり返し計測を含む医療試験における被験者間の相違**

つまり，個体差やブロック間の差といった「推定対象としては関心がないんだけれど，推定に際して無視できないばらつき」があるときに混合モデルを使う，ということだろう．

尤度方程式でみる GLM と GLMM のちがい

一般化線形モデル (GLM)

$$L = \prod_{i\in\{\text{Yes}\}} \frac{1}{1+\exp(-(a+bx_i))} \prod_{i\in\{\text{No}\}} \left(1 - \frac{1}{1+\exp(-(a+bx_i))}\right)$$

$$= \prod_{i\in\{\text{All}\}} \frac{\exp(-(a+bx_i)Y_i)}{1+\exp(-(a+bx_i))}$$

$$Y_i = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if No} \end{cases}$$

一般化線形混合モデル (GLMM)

正規分布がたたみこまれている

$$L = \prod_{i\in\{\text{All}\}} \int_{-\infty}^{\infty} \frac{\exp(-(a_i+bx_i)Y_i)}{1+\exp(-(a_i+bx_i))} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(a_i-\bar{a})^2}{2\sigma^2}\right) \mathrm{d}a_i$$

ここでは「定数項」のパラメーター a_i が正規乱数と仮定

混合モデルを計算する R 関数たち

線形/非線形混合モデル

- `lme()` & `nlme()` (`nlme` library) ……(今日は解説せず)

一般化線形混合モデル (GLMM)

- `glmmPQL()` (in `MASS` library)

  ……罰則つき擬似尤度法による計算，推定にはよいけれどモデル選択とかできん

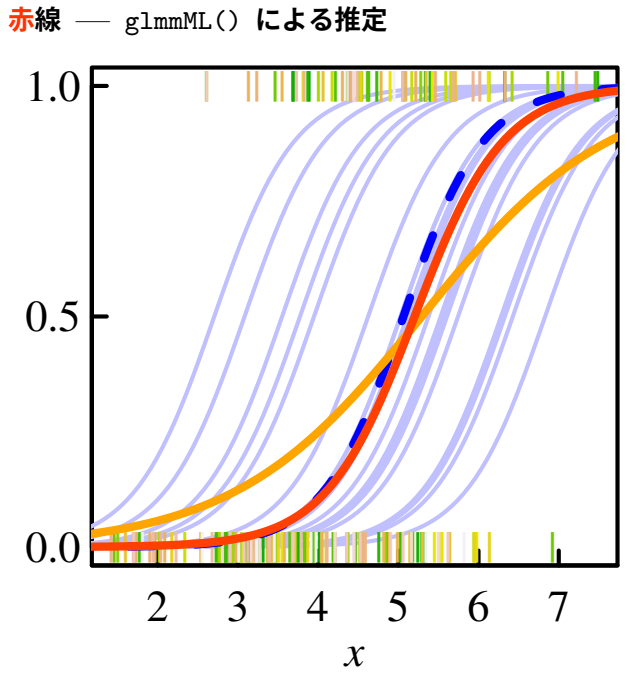
- `glmm()` (in `GLMMGibbs` library)

  ……MCMC 法を使う，2003.12.02 現在，開発中止状態?

- `glmmML()` (in `glmmML` library) ……(今日はこれを使って計算する)
  - 混合モデルの 尤度を数値積分 で計算する
  - 「定数項」 のみがランダム変量になりうる
  - 確率分布 `family = binomial or poisson` だけ
  - まだまだ開発途上 (Göran Broström さんによる)

glmmML() による混合モデルの最尤推定

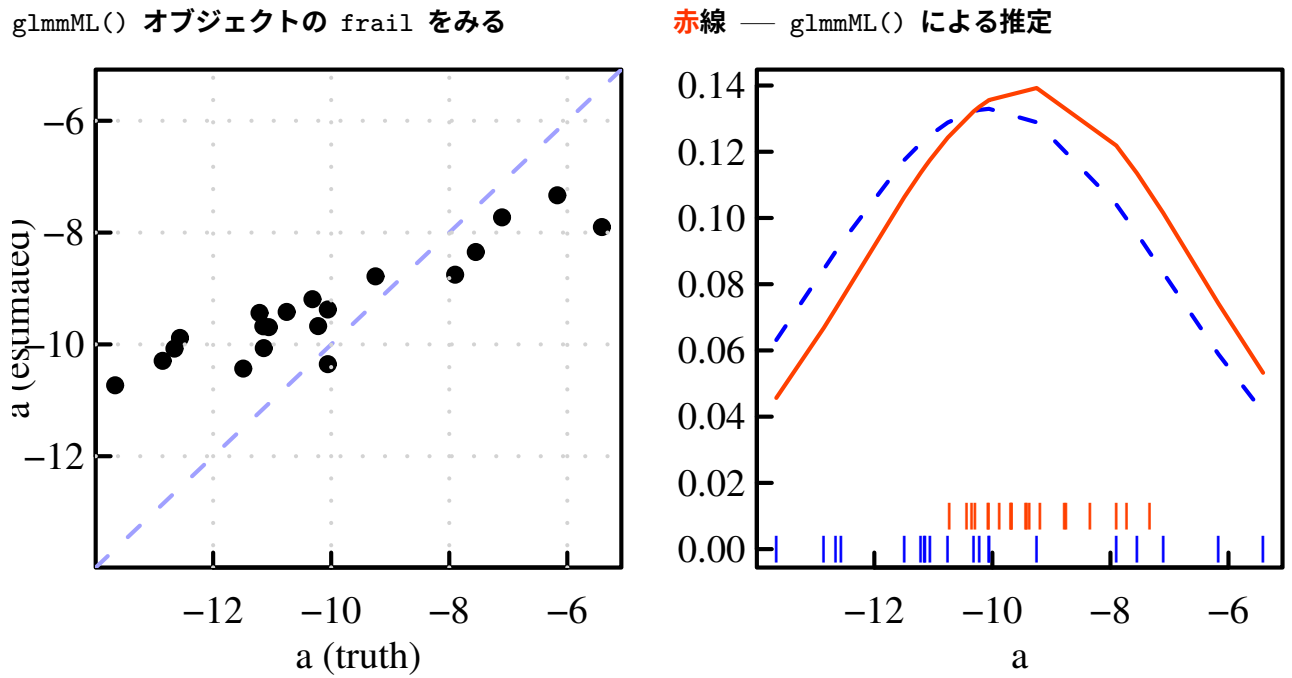
```
glmmML(y ~ 1 + x, cluster = data$id, family = binomial(logit),
       data = data, start.sigma = 3)
```



- glm() より glmmML() はよい推定結果を出す
- 赤池の情報量基準 (AIC) も改善される (a がランダム変量なら)
- 左の数値例はかなりうまくいってる場合

うまく推定できてない $(?)$ 部分

今回の例と似た事例について試した範囲では，個体の「個性」をあらわすパラメーター a の推定に「ずれ」が生じていた



(しかし解釈その他にまちがいあるかも)

glmmML() における AIC によるモデル選択

……は無いので自作しました（たぶん将来だれかがもっとマトモなものを作るはず）

stepAIC.glmmML() の動作例

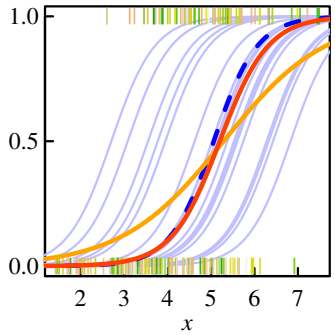
```
Start:  AIC= 3174.27
 dead ~ 1 + date + score + genet + gap + height + distance + density
...(中略)...
#-----------------------------------------------------------------
# STEPWIZE: 2
# formula(AIC = 2735.01): dead ~ 1 + date + score + genet + height + distance
# formula(AIC = 3444.11): dead ~ 1 + score + genet + height + distance
# formula(AIC = 3404.83): dead ~ 1 + date + genet + height + distance
# formula(AIC = 2743.96): dead ~ 1 + date + score + height + distance
# formula(AIC = 2741.59): dead ~ 1 + date + score + genet + distance
# formula(AIC = 2737.76): dead ~ 1 + date + score + genet + height
...(後略)...
```

(来週の大澤君の講座セミナーの内容から借用)

本日のまとめ

1. **データよく見て「個性」の有無をみる**
   - 個体ごとの**ずれ**，overdispersion に注意
2. **「個性」は混合モデルであつかってみる**
   - `nlme()` と `glmmML()` の使い分けに注意
3. **`glmmML()` の推定結果はよく検討してみる**

モデルの予測は観測データをうまく説明できているか，という点検が必要



2003.12.02